Conducting a Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository

Robert R. Downs and Robert S. Chen

Abstract

Long-term preservation and stewardship of scientific data and research-related information is paramount to the future of science and scholarship. Disciplinary and interdisciplinary scientific data archives can offer capabilities for managing and preserving data for research, education, and decision-making activities of future communities representing various scientific and scholarly disciplines. However, meeting the requirements for a trusted digital repository presents challenges to ensure that archived collections will be discoverable, accessible, and usable in the future. Assessing whether scientific data archives meet the requirements for trustworthy repositories will help to ensure that today's collections of scientific data will be available in the future. A continuing self-assessment of a long-term archive for interdisciplinary scientific data is being conducted to identify improvements needed to become a trustworthy repository for managing and providing access to interdisciplinary scientific data by future communities of users. Recommendations are offered for archives of scientific data to meet the requirements of a trustworthy repository.

Scientific Data Stewardship

Today, scientific data are routinely created in digital form and analyzed using computer-based applications. In addition to enabling the creation and analysis of scientific data, the digital form enables data sharing and reuse by others, which has led to the advent of data-driven science practices within scientific communities (Lesk, 2008). Also, online access to scientific data has enabled the development of online learning resources and web services that facilitate uses of the data beyond those envisioned by the original data producers. Given the shelf-life of storage media and the evolution of computer technology, data in digital form presents challenges for organizations, such as scientific data centers, libraries, and government agencies that are responsible for managing and preserving scientific data over time. Many organizations have begun implementing digital repositories to manage their collections of digital information, including scientific data, and to provide long-term stewardship for such collections.

Similar to the intellectual property and digital assets managed by institutional archives and repositories, the content of scientific archives and repositories also need to be trustworthy (Gladney, 2006). If the scientific data and research-related information stored in a digital repository cannot be trusted, then their potential value for use could be questionable. Ensuring the trustworthiness of scientific data held in a digital repository or data archive can help to alleviate concerns of potential data users and data providers. Digital repositories and archives established for the preservation of scientific data must ensure that the data within these facilities maintains its integrity and that the managed data will be preserved to maintain its trustworthiness

for future users, which could include scientists, decision-makers, educators, and learners from the scientific discipline represented by the data. Furthermore, organizations that are responsible for the preservation of scientific data used by multiple disciplines must consider the needs for ensuring the trustworthiness of the data to future groups of potential users representing various disciplines.

Standards, such as the Open Archival Information Systems Framework (CCSDS, 2002), offer guidance for the long-term stewardship of scientific data and other digital resources. Based on such standards for data archiving and digital preservation, instruments have been developed for assessing the trustworthy nature of data archives and other digital repositories. Using these instruments to conduct a self-assessment can assist managers in improving capabilities for long-term stewardship of digital collections. One of the tools developed to assist in the assessment of digital repositories for trustworthiness, is the Trusted Repositories Audit & Certification: Criteria and Checklist (TRAC) document (OCLC and CRL 2007), The TRAC describes the criteria for a trustworthy repository, and is organized into three categories, including “Organizational Infrastructure”, “Digital Object Management”, and “Technologies, Technical Infrastructure, & Security”. The TRAC, offered for use by repositories to conduct a self-assessment of their trustworthiness, is being used to conduct a continuing self-assessment of a long-term archive that was established for the preservation of interdisciplinary scientific data.

Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data

The Socioeconomic Data and Applications Center (SEDAC), operated by the Center for International Earth Science Information Network (CIESIN) of Columbia University, produces, archives, and disseminates scientific data and offers services to improve understanding of human interactions in the environment. Research, decision-making, and educational communities rely on SEDAC to continually provide access to scientific data and services to foster data analysis. After assessing various systems and testing Fedora as a pilot project, CIESIN implemented Fedora, procuring VITAL and maintenance services from VTLS, Inc. Recognizing the need to preserve scientific data that has become less popular, CIESIN established the SEDAC Long-Term Archive (LTA) as a collaborative experiment being conducted by SEDAC, the Columbia University Libraries, and the Earth Institute of Columbia University (Downs, Chen, Lenhardt, Bourne, & Millman, 2006). The LTA is using the TRAC document to conduct the self-assessment in stages, as a continuing exercise, to identify areas for further improvement and to initiate enhancements to maintain its trustworthiness.

Conducting the self-assessment has revealed that the LTA can meet each of the criteria measures if both traditional scientific data management practices and implemented digital repository capabilities are evaluated. However, as the LTA continues to improve its digital repository and adopt other technological enhancements to enhance its data stewardship and preservation

practices, the LTA will need to continue conducting the self-assessment to ensure that the criteria for trustworthiness are still being met as the archival infrastructure and practices evolve.

Three initiatives have contributed to the capabilities of the LTA that could improve the capabilities of other repositories and scientific data archives. These include a strategy for collaborative organizational sustainability, a model for submission of scientific data to the repository, and a plan for facilitating intra-organizational transfer between the repository at CIESIN and the repository managed by the Libraries. These initiatives are introduced as exemplars for archival organizations and repositories to consider as recommendations for initiating improvements for the trustworthiness of their archival policies and procedures, digital repositories, and content of their collections.

Strategy for Collaborative Organizational Sustainability

Collaboratively managing the LTA has fostered a sustainable organizational infrastructure for managing the archive, including its technical infrastructure and the content of its collection. The LTA Board is currently comprised of representatives from the SEDAC, the Columbia University Libraries, and the Earth Institute. The Board determines the appraisal criteria for accession to the LTA and decides whether nominated scientific data sets will be accessioned. In making these decisions, the Board is cognizant of the potential long-term costs of digital preservation. Currently, the LTA is managed by SEDAC staff. In the event that sponsorship discontinues for the operation of the SEDAC, contingency plans change the composition of the Board and the management of the LTA so that the organizational entities can assume control and accept the fiscal responsibilities for continual management and operation of the LTA, if necessary.

Model for Submission of Scientific Data to the Repository

Improving capabilities for producers to submit scientific data to a repository is needed to capture the data and descriptions soon after creation and to improve efficiency. A model has been developed to guide the design of capabilities for web-based submission and workflow for ingest of scientific data to the repository. The model specifies functional capabilities to support data submission services and addresses the TRAC criteria for submission, review, preparation, and ingest. Customization of the open source submission software, VALET, has begun to meet the specifications. User testing should identify additional enhancements that are needed.

Plan for Intra-Organizational Collection Transfer

The Columbia University Libraries are also planning to implement Fedora for long-term digital repository operations. The Libraries and the LTA are developing plans for a series of tests of the capabilities for transferring selected scientific data sets between the two repositories. The results of these tests will assist in the evaluation of the requirements to facilitate future integration of the repository environments.

Challenges for the Future

Various challenges and recommendations for archives and repositories have been identified by engaging in the self-assessment of the LTA. Like the LTA, archives and repositories need to continue self-assessment on a routine basis to ensure that they continue to meet the criteria for trustworthiness and continue to improve the extent to which each criterion is met while completing the transition of infrastructure and collections to digital repository systems. Continuous improvements also need to be applied to archives to ensure that the trustworthiness of data and services are maintained as collections grow, as technologies are adopted, and as new services are offered for current and future user communities. In addition, archives and repositories offering long-term preservation will need to seek certification when standard criteria have been established for external auditors to assess and certify repositories for trustworthiness.

References

Consultative Committee for Space Data Systems (CCSDS). 2002. Reference Model for an Open Archival Information System (OAIS). Adopted as: Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003). Available: http://public.ccsds.org/publications/archive/650x0b1.pdf

Downs, R. R., R. S. Chen, W. C. Lenhardt, W. Bourne, and D. Millman. 2007. Cooperative management of a long-term archive of heterogeneous scientific data. Proceedings of Ensuring the Long-Term Preservation and Value Adding to Scientific and Technical Data (PV 2007). Oberpfaffenhofen/Munich, Germany. October 9–11. Available: http://www.pv2007.dlr.de/Papers/Downs_CooperativeManagementOfALongTermArchive.pdf

Gladney, H. M. 2006. Principles for digital preservation. Communications of the ACM, 49(2), 111 - 116.

Lesk, M. 2008. Recycling Information: Science Through Data Mining. International Journal of Digital Curation, 3(1), 154-157.

OCLC and CRL. 2007. Trustworthy Repositories Audit & Certification: Criteria and Checklist. Available: http://bibpurl.oclc.org/web/16712